Gallery Filter Network for Person Search

Lucas Jaffe and Avideh Zakhor

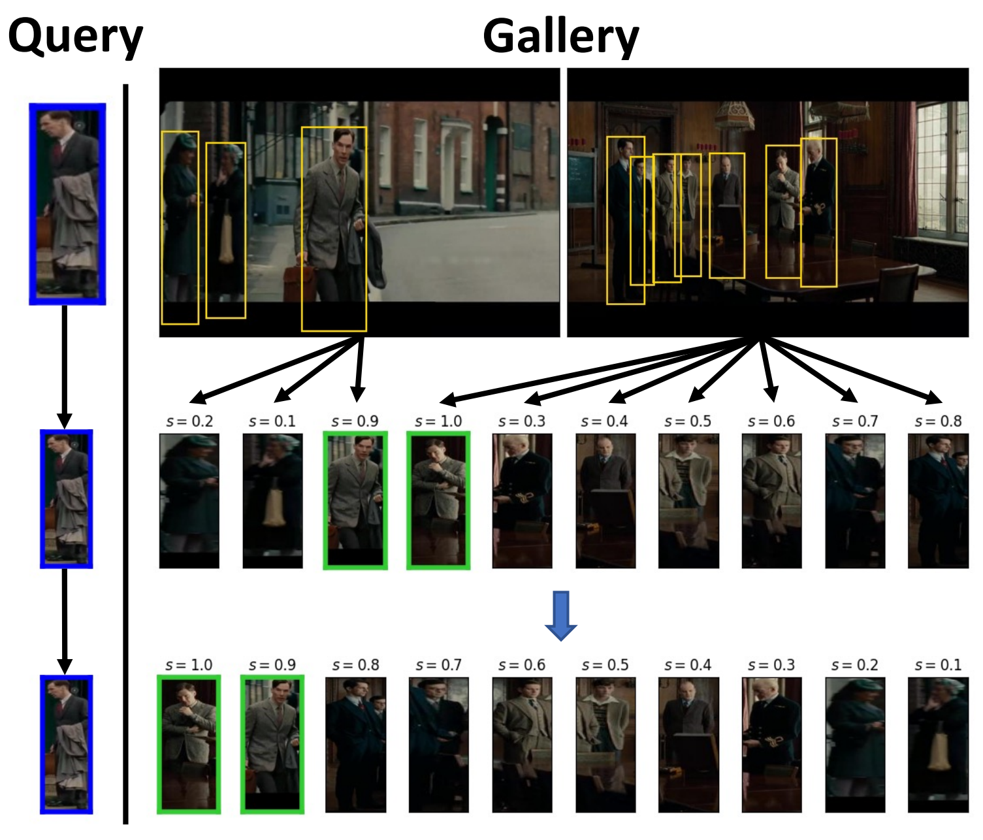
UC Berkeley, EECS

Abstract

In person search, we aim to localize a query person from one scene in other gallery scenes. The cost of this search is dependent on expensive object detection in each gallery scene, making it beneficial to reduce the pool of likely scenes. We propose the Gallery Filter Network (GFN), a novel module which efficiently discards gallery scenes from the search process, and benefits scoring for persons detected in remaining scenes.

1. Background

Person Search: Localize each instance of a *query* person image in a set of scene images called a *gallery*.



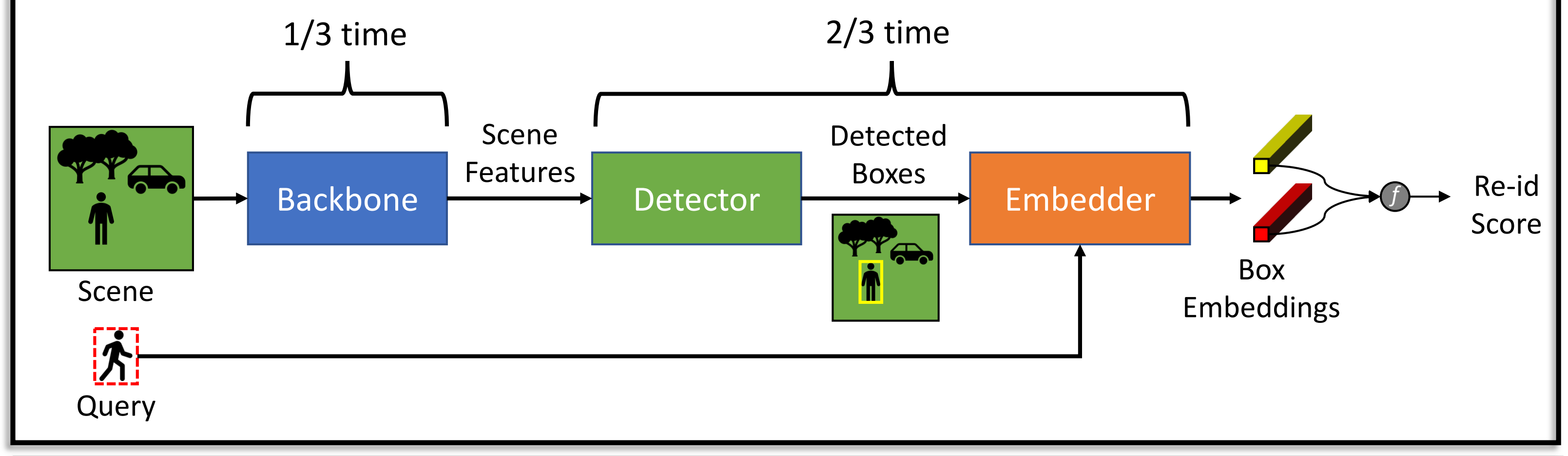
1. **Person Detection:** Localize each person in every *gallery* scene.
2. **Person Retrieval:** Compare a *query* person to each detected person, and sort them by similarity.

Images from *The Imitation Game* (2014), contained in the *CUHK-SYSU* dataset.

2. Problem Statement

Problem: Object detection step of person search is expensive.

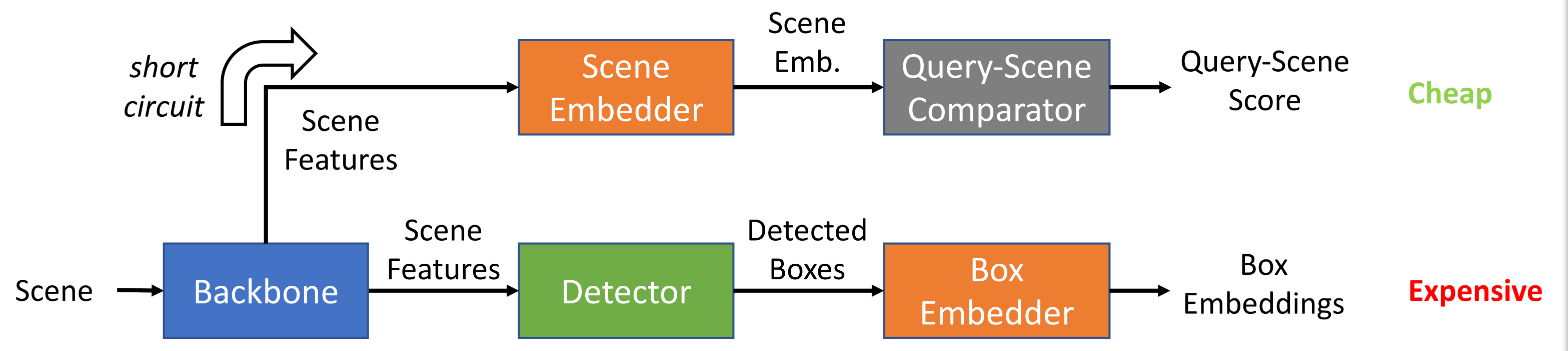
- 1/3 time spent on computing backbone features
- 2/3 time spent on detection and embedding



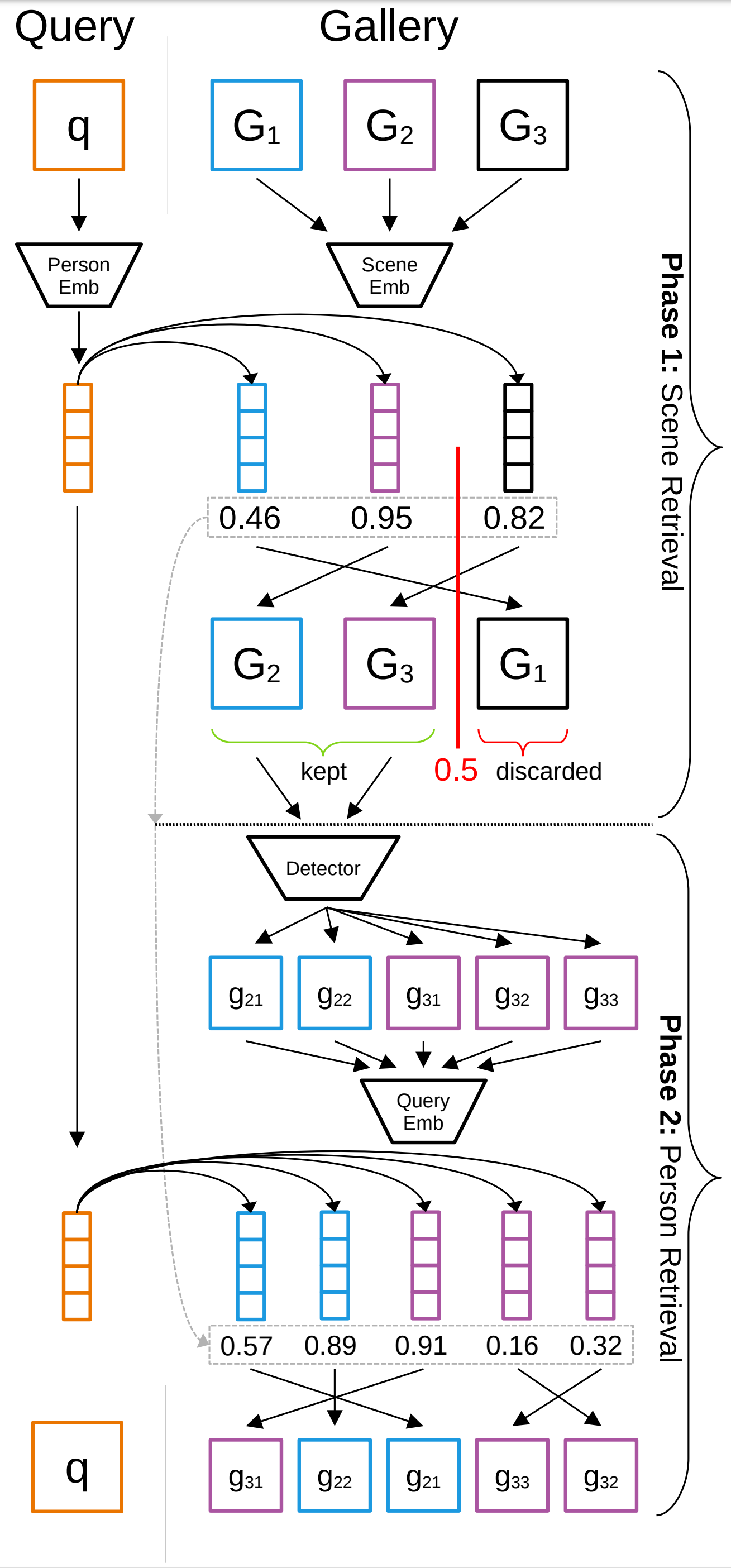
3a. Proposed Method

We propose the Gallery Filter Network (GFN), which avoids detection by splitting person search into two phases:

1. **Scene Retrieval:** Rank scenes by likelihood they contain query person.
2. **Person Retrieval**: Detect and rank persons by similarity to query person.



3b. Model Process



1a. Person, scene embeddings extracted using embedding (Emb) modules.

1b. Query-scene scores computed using cosine similarity of extracted embeddings.

1c. Using hard threshold, low-scoring scenes discarded: no detection needed.

2a. Detection performed on high-scoring scenes.

2b. Embeddings extracted from detected boxes.

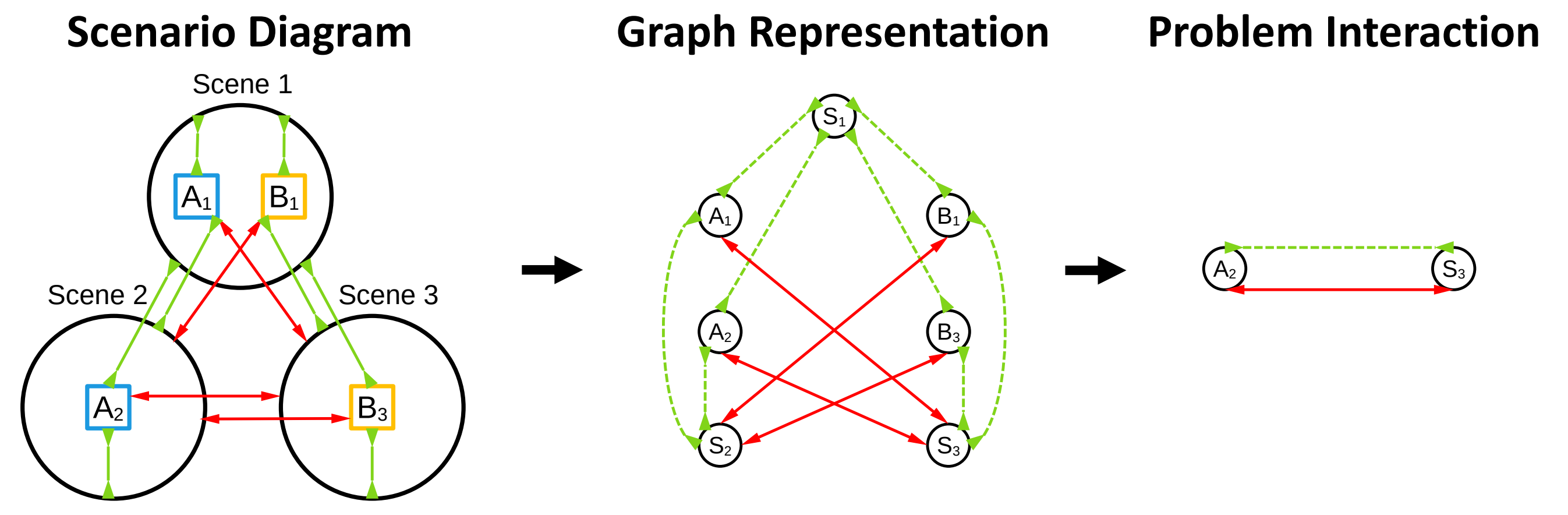
2c. Query-detect similarity scores computed, combined with query-scene scores.

2d. Combined scores sorted to determine final ranking.

3e. Improved Objective

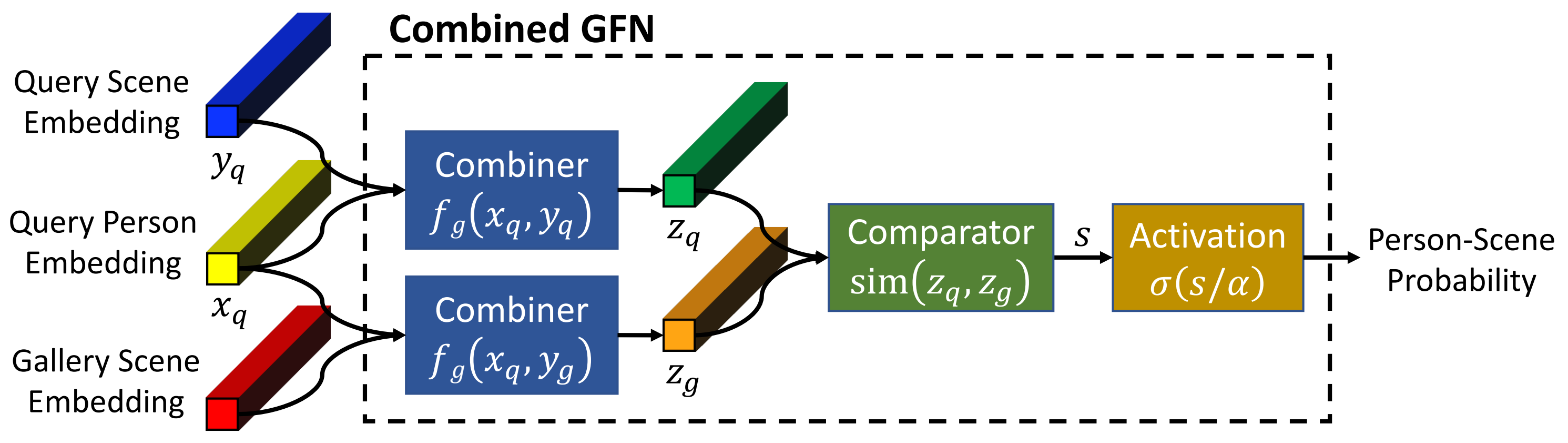
Problem: Baseline system has conflicting attractions and repulsions.

- System below has three scenes, with two person identities (A, B)
- A_2 and S_3 simultaneously pushed together and pulled apart.
- Prevents system from reaching an optimized state.



Solution: Disentangle conflicting interactions by combining person and scene embeddings: $f_g(\vec{x}, \vec{y}) = \mathrm{BN}(\sigma(\vec{x}/\beta) \odot \vec{y})$

$$z_q = f(x_q, y_q), \; z_g = f(x_q, y_g)$$



Objective:
$$\ell_c = -\log \frac{\exp\left(\mathrm{sim}(z_q, z_g^+)/\tau\right)}{\exp\left(\mathrm{sim}(z_q, z_g^+)/\tau\right) + \sum_{z_g^- \in Z_g^-} \exp\left(\mathrm{sim}(z_q, z_g^-)/\tau\right)}$$

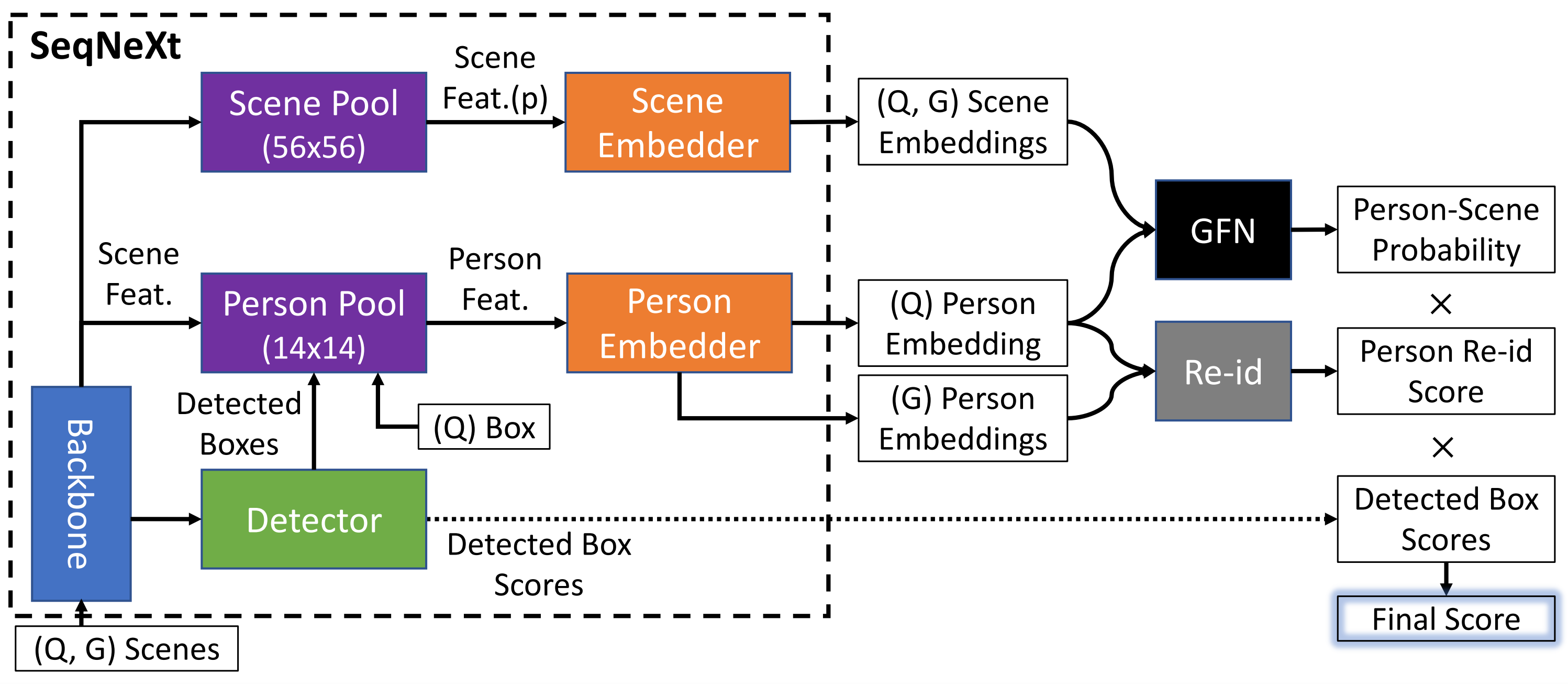
5. Conclusions

The SeqNeXt and GFN models improve person search through:

1. **Efficiency:** GFN is effective for filtering gallery scenes, saving significant compute from detection, embedding.
2. **Accuracy:** SeqNeXt+GFN score weighting improves over SOTA on benchmark datasets for all metrics.
3. **Modularity:** GFN is a modular component which can be appended to any person search model.

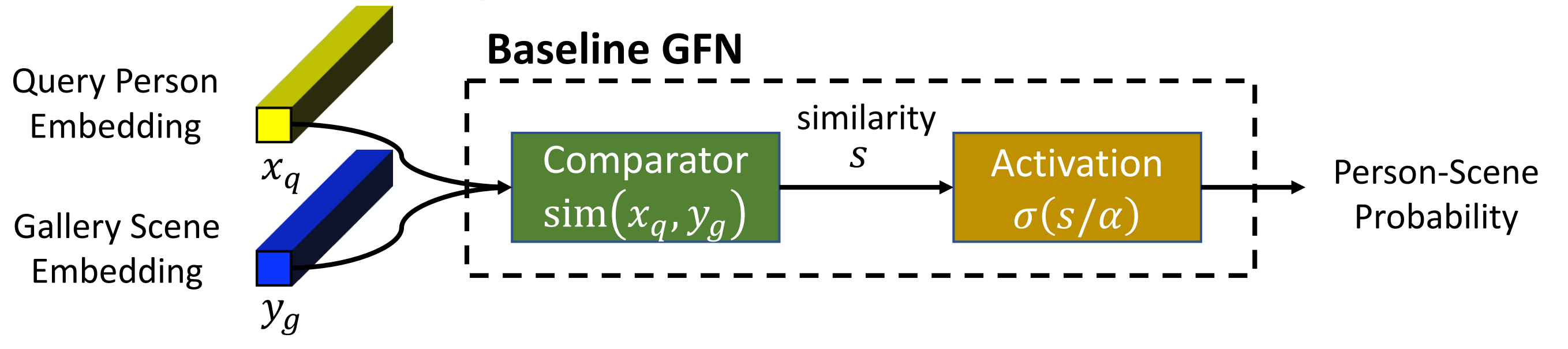
3c. Model Architecture

SeqNeXt: Person search model, improves on previous *SeqNet* [3].



3d. Baseline Objective

GFN Goal: Output a high score when person in scene, low score when not.



Objective: Given person embeddings x_q, scene embeddings y_g:

$$\ell_b = -\log \frac{\exp\left(\mathrm{sim}(x_q, y_g^+)/\tau\right)}{\exp\left(\mathrm{sim}(x_q, y_g^+)/\tau\right) + \sum_{y_g^- \in Y_g^-} \exp\left(\mathrm{sim}(x_q, y_g^-)/\tau\right)}$$

4. Experiments and Analysis

CUHK-SYSU Dataset

PRW Dataset

We conduct experiments on two benchmark datasets: *PRW* [1] and *CUHK-SYSU* [2].

Datasets:

- **CUHK-SYSU:** 18k scenes, 96k persons, 8k identities
- **PRW:** 12k scenes, 43k persons, 1k identities

Training:

- **Optimization:** 30 epochs, Adam, LR=1e-4
- **Augmentation:** 640 × 640 random crops, HFlip
- **Training time:** 20h on CUHK-SYSU, 10h on PRW
- **Hardware:** Quadro RTX 6000 GPU, 24GB VRAM

Re-id Results:

1. SeqNeXt+GFN improves over SOTA *PSTR* [4].
2. GFN score-weighting boosts all metrics.

Method	CUHK-SYSU		PRW	
	mAP	top-1	mAP	top-1
PSTR	95.2	96.2	56.5	89.7
SeqNeXt	96.1	96.5	57.6	89.5
SeqNeXt+GFN	**96.4**	**97.0**	**58.3**	**92.4**

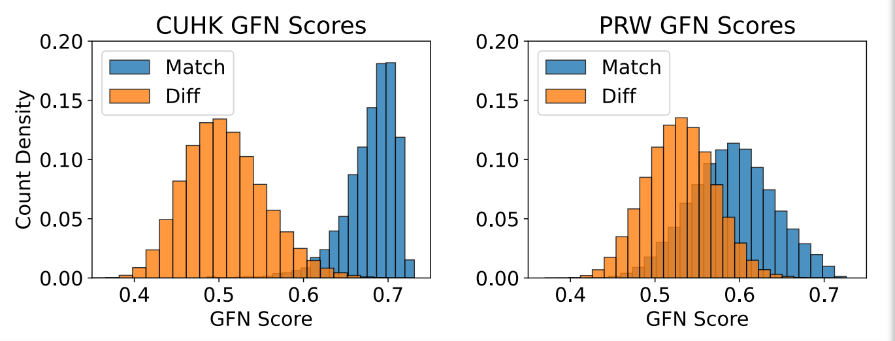
Scene Retrieval Results: GFN effective, but dependent on the dataset: visual diversity of CUHK-SYSU results in greater time saved (TS).

	CUHK-SYSU		PRW	
Recall (%)	95	100	95	100
mAP (%)	91.6	96.4	55.4	58.3
TS (%)	57.8	0.0	13.6	0.0

Compute Time Breakdown:

- 35% backbone, 60% detector, 5% GFN

GFN scores: Histograms reveal CUHK-SYSU clusters more separable.



Acknowledgments

The authors would like to thank Wesam Sakla, Michael Goldman, Omar DeGuchy, and Cindy Gonzales for helpful feedback and discussions.

References

[1] Liang Zheng, Hengheng Zhang, Shaoyan Sun, Manmohan Chandraker, Yi Yang, and Qi Tian. *Person Re-identification in the Wild.* CVPR 2017.

[2] Tong Xiao, Shuang Li, Bochao Wang, Liang Lin, and Xiaogang Wang. *Joint Detection and Identification Feature Learning for Person Search.* CVPR 2017.

[3] Zhengjia Li and Duoqian Miao. *Sequential End-to-end Network for Efficient Person Search.* AAAI 2021.

[4] Jiale Cao, Yanwei Pang, Rao Muhammad Anwer, Hisham Cholakkal, Jin Xie, Mubarak Shah, and Fahad Shahbaz Khan. *PSTR: End-to-End One-Step Person Search With Transformers.* CVPR 2022.